Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:22:38
REFERENCE NO: 281

Author Names & Affiliations

- Elizabeth Brainerd - Brown University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Basic Research, Biology, Organismal Biology

Title of Submission

Video data preservation and sharing for integrative organismal biology research

Abstract (maximum ~200 words).

Cyberinfrastructure for storing and sharing large video files is needed to advance research in integrative organismal biology. These video files are the raw data for research in several fields, such as comparative biomechanics and animal behavior. Substantial time, effort and funding goes into the collection of these scientific videos. Our ability to preserve and share video data is limited by the availability and expense of data storage for video files that are often 10 GB or larger. In the coming years we expect the pace of video data collection to increase, given the advent of low-cost consumer video cameras with high frame rates, such as Go Pro cameras. The base of researchers and students generating these data will broaden to include more small colleges and even K-12 schools. A national infrastructure for hosting video data repositories is needed to collect, preserve and share these valuable raw data. Videos collected for specific scientific purposes in organismal biology are enormously rich and valuable for reuse. Unlike the vast amounts of low quality, downsampled video on sites such as YouTube, scientific video is saved with key metadata such as frame rate, spatial calibration, and scientific name of the species in the video. The future of motion analysis and behavioral classification is machine learning, potentially accelerating the process of video analysis. But we need the cyberinfrastructure to store and share the videos for machine training and data mining.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

A central goal of organismal biology is to quantify complex phenotypes for integration with genotype/expression/genomic data. Such integration holds promise for more sophisticated understandings of embryological development, physiology, ecology, behavior and the evolution of life on earth. Historically, phenotypes such as round, wrinkly, yellow and green were fine for Mendel’s peas. But today we have access to much richer information about genomes and their expression patterns, so more sophisticated phenotype information is also needed. For aspects of phenotype that have a time-dependent component, such as the maximum running speed of a lizard, video imaging

is the best capture medium.

Video imaging is used widely for capturing a range of time-varying phenotypes, from simple motions, such as a frog jump, to complex social behaviors of animals. In some cases, several high-speed video cameras running at 1000 or more frames per second are used to capture rapid 3D motion.

The individual video files from these cameras are large, up to 10 GB, and even a small study can routinely generate a TB of data. Furthermore, we are on the cusp of an explosion in the volume of organismal video, given the advent of low-cost consumer video cameras with high frame rates, such as Go Pro cameras. Until recently, high-speed cameras cost several tens of thousands of dollars and were found only in high-end research labs. But consumer cameras are accessible to researchers and students at small colleges and even K-12 schools, broadening research opportunities in organismal biology.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Organismal biology needs infrastructure to support data repositories for storing and sharing large video files (typically up to 10 GB at the moment; will be larger in the future). Existing data repositories limit the size of individual files, and often the durations of uploads and downloads are impractically long. Commercial data storage is expensive and does not meet the need for archival preservation of video data as would be the goal of dedicated repositories for scientific video.

The specific need is for (1) computing infrastructure for hosting VMs to run the back end databases and front-end web interfaces for video data and metadata deposition and sharing, and (2) large capacity storage without file-size limitations for storing video data. The storage system must meet digital data preservation standards for protection of the data.

Video data collected for specific scientific purposes in organismal biology are enormously rich and valuable for reuse. Unlike the vast amounts of low quality, downsampled video on sites such as YouTube, scientific video is saved with key metadata such as frame rate, spatial calibration, and scientific name of the species in the video. The future of motion analysis and behavioral classification is machine learning, potentially vastly speeding up the process of video analysis. But we need the cyberinfrastructure to store and share the videos for machine training and data mining.

Consent Statement

- “I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).”